DOCUMENT RESUME

ED 292 832 TM 011 210

AUTHOR Holland, Paul W.
TITLE The Dutch Identity: A New Tool for the Study of Item Response Theory Models. Program Statistics Research. Technical Report No. 87-78. Research Report No. 87-27.
INSTITUTION Educational Testing Service, Princeton, NJ. Program Statistics Research Project.
PUB DATE Jun 87
NOTE 27p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Factor Analysis; *Latent Trait Theory; Models
IDENTIFIERS *Dutch Identity (Testing); Latent Class Models; Linear Models; Log Linear Models

ABSTRACT

The Dutch Identity is a useful way to reexpress the basic equations of item response theory (IRT) that relate the manifest probabilities to the item response functions (IRFs) and the latent trait distribution. The identity may be exploited in several ways. For example: (1) to show how IRT models behave for large numbers of items--they are submodels of second-order, log-linear models for 2(sup J) tables; (2) to suggest new ways to assess the dimensionality of IRT models--factor analysis of matrices composed of second-order interactions from log-linear models; (3) to give insight into the structure of latent class models; and (4) to illuminate the problem of identifying the IRFs and the latent trait distribution from sample data. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-87-27

The Dutch Identity: A New Tool for the Study of Item Response Theory Models

Paul W. Holland

ED292832

PROGRAM STATISTICS RESEARCH

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Helen C. Weidenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TECHNICAL REPORT NO. 87-78

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541

TM 011 210

BEST COPY AVAILABLE

THE DUTCH IDENTITY: A NEW TOOL FOR THE STUDY OF ITEM RESPONSE THEORY MODELS

Paul W. Holland

Program Statistics Research
Technical Report No. 87-78

Research Report No. 87-27

Educational Testing Service
Princeton, New Jersey 08541-0001

June 1987

Copyright © 1987 by Educational Testing Service. All rights reserved.

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants. Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

TABLE OF CONTENTS

		Page
1.	Notation	1
2.	The Dutch Identity	4
3.	Some Applications of the Dutch Identity	6
	3.1 An IRT Model That Is A Second-order Log-linear Model	6
	3.2 IRT Models With Large Numbers of Items	9
	3.3 The Study of Test Dimensionality	11
	3.4 Latent Class Models	12
	3.5 What Does An Observed Response Vector Tell Us About the Value of A Latent Variable? ..	14
	3.6 The Rasch Model	16
4.	Discussion	18
	References	20

ABSTRACT

The Dutch Identity is a useful way to reexpress the basic equations of item response theory (IRT) that relate the manifest probabilities to the item response functions (IRFs) and the latent trait distribution. The identity may be exploited in several ways. For example: (a) to show how IRT models behave for large numbers of items -- they are submodels of second-order log-linear models for 2^J tables; (b) to suggest new ways to assess the dimensionality of IRT models -- factor analysis of matrices composed of second-order interactions from log-linear models; (c) to give insight into the structure of latent class models; and (d) to illuminate the problem of identifying the IRFs and the latent trait distribution from sample data.

There are few mathematical tools that have proved useful in the study of the structure of item response theory (IRT) models. This is especially true for the so-called "marginal maximum likelihood" approach in which the distribution of the latent variable is integrated out and the would-be analyser is left facing an intractable integral that must be evaluated numerically (Bock and Lieberman, 1970). While the EM algorithm (Bock and Aitken, 1981) can be used to simplify this integration, this fact is mainly useful in computing maximum likelihood estimates and does not lead to any insight into the structure of the models themselves.

The purpose of this paper is to introduce a new tool that does, in some cases, make the integrals disappear and allows the structure of the model to appear in useful new ways. The remainder of this paper is organized as follows. Section 1 sets up the notation; section 2 states and proves the new result -- the Dutch Identity. Section 3 illustrates its value in several problems and section 4 contains additional discussion.

1. NOTATION

The notation follows that in Holland (1981), Cressie and Holland (1983), and Holland and Rosenbaum (1986). We let C denote a population of examinees and T a specific test. The zero-one variable x_j denotes correct or incorrect on item j in T and the response vector, $\mathbf{x}$, is given by:

$$\mathbf{x} = (x_1, \ldots, x_J).$$

Let the proportion of examinees in the population, C, who would produce response vector $\mathbf{x}$ when tested with T be denoted $p(\mathbf{x})$. Clearly, we have

$$p(\mathbf{x}) \geq 0 \quad \text{and} \quad \sum_{\mathbf{x}} p(\mathbf{x}) = 1.$$

The 2^J values, $p(\mathbf{x})$, are called the manifest probabilities. Let X be the response vector of a randomly selected examinee from C on test T. The probability function for X, $\text{Prob}\{X=\mathbf{x}\}$, is just $p(\mathbf{x})$, i.e.

$$\text{Prob}\{X=\mathbf{x}\} = p(\mathbf{x}).$$

Item response models restrict the form of the manifest probabilities, $p(\mathbf{x})$, in the following way. First of all, the value of a latent variable, θ, is assumed to be associated with each examinee in C such that

a) given θ, the coordinates of X are independent, i.e.

$$P(X=\mathbf{x}|\theta) = \prod_j P(X_j=x_j|\theta), \quad \text{and}$$

b) the item response functions, $P(X_j=1|\theta) = P_j(\theta)$, are usually restricted in some way, e.g., to be monotone increasing in θ or to have a specified functional form such as the one-, two-, or three-parameter logistic form (Birnbaum, 1968), and

c) the distribution function of θ over C is denoted by $F(\theta)$.

Since x_j is dichotomous, we may write

$$P(X_j=x_j|\theta) = P_j(\theta)^{x_j}\, Q_j(\theta)^{1-x_j},$$

where $Q_j(\theta) = 1-P_j(\theta)$. The conditional independence assumption may then be written as

$$P(X=\mathbf{x}|\theta) = \prod_j P_j(\theta)^{x_j}\, Q_j(\theta)^{1-x_j}.$$

But, by the usual rules for manipulating conditional probabilities, we have

$$P(X=\mathbf{x}) = \int P(X=\mathbf{x}|\theta)\, dF(\theta),$$

and consequently all (locally independent) item response models may be viewed as

restricting $p(\mathbf{x})$ to have the form

$$p(\mathbf{x}) = \int \prod_j P_j(\theta)^{x_j} Q_j(\theta)^{1-x_j} \, dF(\theta). \qquad (1)$$

Equation (1) relates the manifest probabilities, $\{p(\mathbf{x})\}$, to the latent item response functions, $\{P_j(\theta)\}$, and the distribution of the latent variable, $F(\theta)$. In this paper, as in the discussion of the marginal maximum likelihood approach, (1) is taken to be the defining characteristic of any IRT model. The integral in (1) is the "intractable integral" referred to earlier and is often an obstacle to the further understanding of IRT models.

The manifest probabilities, $\{p(\mathbf{x})\}$, are the governing quantities in the likelihood function that arises when data are collected by randomly sampling N examinees from C and testing them with T. In this situation let

$n(\mathbf{x})$ = number of examinees in the sample producing response vector $\mathbf{x}$.

Then, if C is large compared to N, $\{n(\mathbf{x})\}$ follows a multinomial distribution with parameters N and $\{p(\mathbf{x})\}$. The likelihood function is the multinomial probability function (except for a multiplicative constant) and given by

$$\prod_{\mathbf{x}} p(\mathbf{x})^{n(\mathbf{x})} .$$

Thus the log-likelihood function is

$$L = \sum_{\mathbf{x}} n(\mathbf{x}) \log p(\mathbf{x}). \qquad (2)$$

Hence, it is natural to study the structure of $\log p(\mathbf{x})$, i.e., the "log-manifest probabilities," and I shall do just that.

In this context, a <u>model</u> for $p(\mathbf{x})$ is simply a <u>restriction</u> on the form of $p(\mathbf{x})$ in (2) and, in particular, IRT models are formed by equation (1) and possible restrictions on the $\{P_j(\theta)\}$ and $F(\theta)$.

Cressie and Holland (1983) studied the structure of the models defined by (1) and were successful in completely characterizing p(x) in the case of the Rasch model -- the case where the IRFs have the form specified by

$$\log\left(\frac{P_j(\theta)}{Q_j(\theta)}\right) = a(\theta - b_j). \tag{3}$$

In (3), a = common discrimination parameter, and b_j is the item difficulty parameter. In this paper, I will generalize a formula obtained by Cressie and Holland that re-expresses (1) in a useful way. This generalization is the Dutch Identity.*

2. THE DUTCH IDENTITY

Theorem 1 gives the basic result of this paper.

Theorem 1: (The Dutch Identity) If p(x) satisfies (1) then for any zero-one vector y

$$\frac{p(x)}{p(y)} = E\left(\exp\left\{\sum_j (x_j - y_j)\lambda_i\right\} \middle| X=y\right) \tag{4}$$

where $\lambda_j = \lambda_j(\theta)$ is the item logit function,

$$\lambda_j(\theta) = \log\left(\frac{P_j(\theta)}{Q_j(\theta)}\right).$$

Before going through the easy proof of (4), let me make a few comments about it. First of all, in (4), x is thought of as varying over all possible response vectors while y is thought of as a fixed response vector. In a sense y is an arbitrary choice of "origin." In using this identity we may choose y to

---

* I discovered this identity and some of its consequences while lecturing in the Netherlands during October, 1986. Since this discovery was in no small part due to the stimulating psychometric atmosphere in Holland, I decided to call the result the Dutch Identity.

have desirable properties. The right-hand-side of (4) is a conditional expectation of a certain function that involves, $\lambda(\theta) = (\lambda_1(\theta), \ldots, \lambda_J(\theta))$, given that X=y. More specifically, it is the posterior moment-generating-function of $\lambda = \lambda(\theta)$ given X=y, evaluated at the point, x-y.

Proof: We may express (1) as

$$p(x) = \int \prod_j \left(\frac{P_j(\theta)}{Q_j(\theta)}\right)^{x_j} \prod_j Q_j(\theta)\, dF(\theta)$$

$$= \int \exp\{\sum_j (x_j\, \lambda_j(\theta)\} \prod_j Q_j(\theta)\, dF(\theta)$$

$$= \int \exp\{\sum_j (x_j - y_j)\, \lambda_j(\theta)\} \prod_j \left(\frac{P_j(\theta)}{Q_j(\theta)}\right)^{y_j} \prod_j Q_j(\theta)\, dF(\theta)$$

so

$$p(x) = \int \exp\{\sum_j (x_j - y_j)\, \lambda_j(\theta)\} \prod_j P_j(\theta)^{y_j}\, Q_j(\theta)^{1-y_j}\, dF(\theta)$$

and

$$\frac{p(x)}{p(y)} = \int \exp\{\sum_j (x_j - y_j)\, \lambda_j(\theta)\} \left[\prod_j \frac{P_j(\theta)^{y_j}\, Q_j(\theta)^{1-y_j}}{p(y)}\, dF(\theta)\right]$$

However, the quantity in brackets is the posterior distribution function of θ given X=y, i.e.,

$$\prod_j \frac{P_j(\theta)^{y_j}\, Q_j(\theta)^{1-y_j}\, dF(\theta)}{p(y)} = dF(\theta \mid X=y)$$

Thus

$$\frac{p(x)}{p(y)} = \int \exp\{\sum_j (x_j - y_j)\, \lambda_j(\theta)\}\, dF(\theta \mid X=y)$$

$$= E(\exp\{\sum_j (x_j - y_j)\lambda_j\} \mid X=y) \qquad \text{QED.}$$

This proof follows the type of argument used by Cressie (1982) to prove a similar type of identity that is useful in empirical Bayes applications. To my knowledge, the Dutch Identity has never been used in the analysis of IRT models, although Cressie and Holland (1983) derived the special case of (4) in which y=0. Finally, it should be mentioned that in (4) the fact that θ is a scalar is not used and in fact θ might be a vector, $\boldsymbol{\theta}$.

3. SOME APPLICATIONS OF THE DUTCH IDENTITY

3.1 An IRT Model That Is A Second-order Log-linear Model

An IRT model for $p(\mathbf{x})$ involves an integral, but log-linear models for $p(\mathbf{x})$ are much simpler and merely state that $\log p(\mathbf{x})$ is linear in some parameters, i.e.

$$\log p(\mathbf{x}) = \alpha + b(\mathbf{x})\,\boldsymbol{\beta} \tag{5}$$

where $\boldsymbol{\beta}$ is a (column) vector of free parameters of length K, $b(\mathbf{x})$ is a (row) vector of K known constants, and α is the normalizing constant that insures that the $p(\mathbf{x})$ sum to 1. Log-linear models for $p(\mathbf{x})$ correspond to log-linear models for 2^J-contingency tables. These are widely used (e.g., Bishop, Fienberg, and Holland, 1975). Some examples are as follows. Throughout the rest of this paper, t denotes vector or matrix transpose.

a) Independence. The coordinates of $X = (X_1,\ldots,X_J)$ are independent if and only if

$$\log p(\mathbf{x}) = \alpha + \sum_j \beta_j x_j. \tag{6}$$

In this case $b(\mathbf{x}) = (x_1,\ldots,x_J)$ and, $\boldsymbol{\beta}^t = (\beta_1,\ldots,\beta_J)$.

b) Generalized Rasch Model. In Cressie and Holland (1983) the following model is discussed in detail

$$\log p(\mathbf{x}) = \alpha + \sum_j \beta_j x_j + \sum_k \gamma_k\, \delta(k,x_+). \tag{7}$$

where $\delta(k,x_+) = \begin{cases} 1 & \text{if } x_+ = k \\ 0 & \text{otherwise,} \end{cases}$ and $x_+ = \sum_j x_j$.

If $b(x) = (x_1,\ldots,x_J,\ \delta(1,x_+),\ \ldots\ \delta(J,x_+))$

and $\beta^t = (\beta_1,\ldots,\beta_J,\ \gamma_1,\ldots,\gamma_J)$ then (7) defines the class of extended Rasch models. If the $\{\gamma_k\}$ are restricted by the inequalities indicated by Cressie and Holland (1983), then (7) defines the class of Rasch models.

c) **Second-order Exponential Models**. Tsao (1967) defines a second-order exponential (SOE) model by

$$\log p(x) = \alpha + \sum_j \beta_j x_j + \sum_{r<s} \gamma_{rs}\, x_r x_s. \qquad (8)$$

In this case $b(x) = (x_1,\ldots,x_J,\ x_1x_2,x_1x_3,\ldots,x_{J-1}x_J)$,

and $\beta^t = (\beta_1,\ldots,\beta_J,\ \gamma_{12},\ \gamma_{13},\ldots,\gamma_{J-1,J})$.

An interesting question is whether or not an IRT model satisfying (1) can ever be equivalent to a SOE model. This section shows that from the Dutch Identity one may construct an IRT model that is a submodel of the class of SOE models. The next section shows that this construction is far more general than it might first appear. I will state the results as a corollary to Theorem 1 in which θ is a column vector.

Corollary 1: If, for some choice of y, the posterior distribution of $\theta|X=y$ is normal, i.e.

$$F(\theta|X=y) \text{ is } N_D(\mu_y,\ \textstyle\sum_y),$$

and if the item logit functions $\lambda_j(\theta)$ are linear, i.e.

$$\lambda_j(\theta) = \lambda_j(\mu_y) + a_j(\theta-\mu_y)$$

where $a_j = (a_{1j},\ldots,a_{Dj})$ and D is the dimensionality of θ then

$$\log p(x) = \alpha + (x-y)^t\, \lambda(\mu_y) + \tfrac{1}{2}(x-y)^t\, A \textstyle\sum_y A^t(x-y) \qquad (9)$$

where $A^t = (a_1^t, \ldots, a_J^t)$, and $\alpha = \log p(y)$.

I first prove this result using Theorem 1 and then I will comment on it.

Proof: From (4) we have

$$\log p(x) = \alpha + \log E(e^{(x-y)^t\lambda} \mid X=y), \tag{10}$$

where $\alpha = \log p(y)$. But by assumption $\theta|X=y$ is $N(\mu_y, \sum_y)$, and since $\lambda(\theta)$ is a linear function of θ, the posterior distribution of λ is also multivariate normal. Hence, we have

$$E(\lambda|X=y) = \lambda(\mu_y)$$

and (11)

$$\mathrm{Cov}(\lambda|X=y) = A \sum_y A^t.$$

Now remember that the expected value in (10) is the moment generating function (mgf) of λ evaluated at $(x-y)$. However, the mgf of a normal variable Z with mean μ and covariance $\sum$ evaluated at s is

$$E[\exp\{s^tZ\}] = \exp\{s^t\mu + \tfrac{1}{2} s^t \sum s\}. \tag{12}$$

Applying (12) to (11) and (10) with $s = x-y$ and taking logs yields (9). QED.

To see that (9) is, in fact, of the form (8), expand the terms in (9) and collect them to form

$$\log p(x) = \{\alpha + \tfrac{1}{2} y^t B y - y^t \lambda(\mu_y)\} + x^t\{\lambda(\mu_y) - B y\} + \tfrac{1}{2} x^t B x, \tag{13}$$

where $B = A \sum_y A^t$. Now suppose $B = \Gamma + D_b$ where Γ has a zero diagonal, b is the diagonal of B and D_b is the diagonal matrix based on b. Thus (9) is equivalent to

$$\log p(x) = \{\alpha + \tfrac{1}{2} y^t B y - y^t \lambda(\mu_y)\} + x^t\{\lambda(\mu_y) - B y + b\} + \tfrac{1}{2} x^t \Gamma x, \tag{14}$$

since $x_i^2 = x_i$.

If we now make the substitution

$$\alpha' = \alpha + \tfrac{1}{2} y^t B y - y^t \lambda(\mu_y)$$

$$\beta = \lambda(\mu_y) - B\, y + b,$$

we see that (9) is equivalent to

$$\log p(x) = \alpha' + x^t \beta + \tfrac{1}{2} x^t \Gamma x, \tag{15}$$

which is just a matrix way of expressing (8).

The fact that an IRT model can exist that is a non-trivial example of a SOE model (i.e., is not independent) is quite interesting in its own right. Lord (1962) showed that second-order linear (as opposed to log-linear) models do not give reasonable score distributions in general. This would not be true of the model specified in (9) or (15).

3.2 IRT Models With Large Numbers of Items

It might be thought that the example given in Corollary 1 is unusual but the purpose of this section is to show that it holds as a limiting form for all "smooth" unidimensional IRT models.

When the number of items, J, is large, θ is a scalar, F has a density, and y is a "typical" response vector, then the posterior distribution of θ given X=y is approximately normal, i.e.

$$dF(\theta|X=y) = \frac{1}{\sigma_y}\, \phi\left(\frac{\theta-\mu_y}{\sigma_y}\right)\, d\theta, \tag{16}$$

where $\phi(x)$ is the unit normal density function. Furthermore if the item logit functions, $\lambda_j(\theta)$, are differentiable they have the expansion

$$\lambda_j(\theta) = \lambda_j(\mu_y) + \left(\frac{\partial \lambda_j}{\partial \theta}\right)(\theta-\mu_y) + o(|\theta-\mu_y|). \tag{17}$$

Finally, if σ_y is small, as it will be for large enough J, the higher order terms in (17) can be ignored and we have λ approximately multivariate normal with mean vector, $\lambda(\mu_y)$, and covariance matrix, $\left(\frac{\partial \lambda}{\partial \theta}\right) \sigma_y^2 \left(\frac{\partial \lambda}{\partial \theta}\right)^t$, a J×J matrix of

rank 1. Hence, because of Corollary 1, in this situation the following equation will hold approximately (as $J \to \infty$) for any unidimensional IRT model for which the IRFs are smooth, and F is continuous:

$$\log p(x) = \alpha + (x-y)^t \lambda(\mu_y) + \tfrac{1}{2} \sigma_y^2 (x-y)^t \left(\frac{\partial \lambda}{\partial \theta}\right)\left(\frac{\partial \lambda}{\partial \theta}\right)^t (x-y). \qquad (18)$$

Equation (18) defines a submodel of the class of SOE models in (8) in which the second-order parameters are restricted to a multiplicative form. In terms of the free parameters that can be independently estimated, (18) is of the following log-quadratic form:

$$\log p(x) = \alpha + (x^t \beta) + (x^t \gamma)^2. \qquad (19)$$

Equation (19) does not define a log-linear model but rather a submodel of the class of SOE models that has only 2J-parameters rather than the full set of $J + \binom{J}{2}$ parameters of the general SOE model.

The derivation of (18) depends only on the fact that (a) θ is one-dimensional, (b) F is continuous, (c) J is large, (d) y is chosen so that $F(\theta|X=y)$ is approximately normal with a small variance σ_y^2, and (e) $\lambda_j(\theta)$ is differentiable. Since all models in use usually assume (a), (b), and (e) and since the existence of y satisfying (d) is well-known among users of BILOG (see, for example, Bock and Mislevy, 1982), the representation of p(x) as a model of the form (18) is a very general result. The only issue is how large J needs to be for it to hold. This is a worthy topic for future research.

One implication of (18) is that there can be at most two parameters per item consistently estimated for long tests. This is in accord with the general fact that it is difficult to estimate three or more item parameters in an unrestricted fashion for data sets that involve many examinees and many items,

even though IRFs are often parameterized with more than two parameters.

3.3 The Study of Test Dimensionality

If θ is a vector parameter, $\boldsymbol{\theta}$, and has an approximate normal posterior distribution $F(\boldsymbol{\theta}|X=y)$ for some y, with mean μ_y and covariance matrix Σ_y then the generalization of (18) is

$$\log p(x) = \alpha + (x-y)^t \lambda(\mu_y) + \tfrac{1}{2}(x-y)^t \left(\frac{\partial \lambda}{\partial \boldsymbol{\theta}}\right) \Sigma_y \left(\frac{\partial \lambda}{\partial \boldsymbol{\theta}}\right)^t (x-y). \qquad (20)$$

Letting

$$\beta_j = \lambda_j(\mu_y) \quad \text{and} \quad R = \left(\frac{\partial \lambda}{\partial \boldsymbol{\theta}}\right) \Sigma_y \left(\frac{\partial \lambda}{\partial \boldsymbol{\theta}}\right)^t$$

we have

$$\log p(x) = \alpha + (x-y)^t \beta + \tfrac{1}{2}(x-y)^t R(x-y). \qquad (21)$$

Equation (21) says that $p(x)$ satisfies an SOE model in which the matrix of second-order interactions is proportional to R. However, the rank of R is the rank of Σ_y which is the same as the dimensionality of the latent variable $\boldsymbol{\theta}$. Hence, equation (21) suggests that a way to factor-analyze dichotomous items is to fit a SOE model to the 2^J table, $\{n(x)\}$, and then to factor-analyze the matrix of second-order interactions, R. This method will be especially appropriate when there are large numbers of items. It does not make any assumption other than those made in section 3.2.

The matrix of second-order interactions in a SOE (or log-linear) model is only a triangular array with no meaningful diagonal. Hence "factor analysis" of such data is not easily interpretable in terms of covariance matrices and linear regressions of items on factor scores. Instead, all I mean by factor analysis is the decomposition of the elements γ_{rs} in (8) into the following terms

$$\gamma_{rs} = \gamma_r^{(1)} \gamma_s^{(1)t} + \gamma_r^{(2)} \gamma_s^{(2)t} + \ldots \gamma_r^{(D)} \gamma_s^{(D)t} \quad \text{for } 1 \le r < s \le J, \text{ and } D \ll J. \qquad (22)$$

The vectors $\gamma^{(k)} = (\gamma_1^{(k)}, \ldots, \gamma_J^{(k)})^t$ must also satisfy orthogonality constraints. The lengths of these vectors must also decrease, $\|\gamma^{(1)}\| \geq \|\gamma^{(2)}\| \geq \ldots$.

In a more general vein, I am tempted to propose that a test measures D dimensions in population C if representation (1) holds for its manifest probabilities in population C with $\theta = (\theta_1, \ldots, \theta_D)$ and if there is a response vector y such that $F(\theta|X=y)$ is more concentrated about its center in every direction than $F(\theta)$ is. If $F(\theta)$ and $F(\theta|X=y)$ both possess covariance matrices, Σ and Σ_y, then this condition could be expressed as

$$\Sigma - \Sigma_y > 0 \tag{23}$$

in the sense that this difference can be positive definite. This proposal is based on the idea that if the test really measures <u>all</u> of the coordinates of θ then, for at least one response vector, y, our knowledge of θ ought to be more precise <u>in every</u> θ-direction if the response y is observed than if the test is not given, in which case all that is available is the unconditional distribution of θ.

3.4 <u>Latent Class Models</u>

The simplest latent class model has just two latent classes, which we can label by two real numbers θ_1 and θ_2. Then equation (1) reduces to

$$p(x) = \sum_i \{\prod_j P_j(\theta_i)^{x_j} Q_j(\theta_i)^{1-x_j}\} p_i \tag{24}$$

where $p_1 + p_2 = 1$ are the proportions of examinees in C with θ_1 and θ_2, respectively.

This latent class model violates the assumption that F is continuous in the strongest possible way, i.e., F is a two-point distribution. However, the Dutch Identity, (4), is still valid for this case. The posterior distribution $F(\theta|X=y)$

is also a two-point distribution concentrated on θ_1 and θ_2 with

$$p_1(y) = P(\theta=\theta_1 | X=y)$$

and

$$p_2(y) = P(\theta=\theta_2 | X=y).$$

Hence the moment generating function in (4) is given by

$$E(\exp\{\Sigma_j(x_j-y_j)\lambda_j\}|X=y) = p_1(y)\exp\{\Sigma_j(x_j-y_j)\lambda_{j1}\} + p_2(y)\exp\{\Sigma_j(x_j-y_j)\lambda_{j2}\}$$

where $\lambda_{j1} = \lambda_j(\theta_1)$, $\lambda_{j2} = \lambda_j(\theta_2)$. Applying the Dutch Identity yields

$$\frac{p(x)}{p(y)} = p_1(y)\exp\{\Sigma(x_j-y_j)\lambda_{j1}\}\left[1 + \frac{p_2(y)}{p_1(y)}\exp\{\Sigma(x_j-y_j)(\lambda_{j2}-\lambda_{j1})\}\right].$$

Let $\delta_j = \lambda_{j2} - \lambda_{j1}$ and $\beta_j = \lambda_{j1}$, then taking logs we have

$$\log p(x) = \alpha + \Sigma_j(x_j-y_j)\beta_j + \log(1 + e^{\gamma + \Sigma_j(x_j-y_j)\delta_j}) \qquad (25)$$

where $\alpha = \log(p(y)p_1(y))$ and $\gamma = \log(p_2(y)/p_1(y))$.

Let LP(x) be the "logistic potential" function, i.e.

$$LP(x) = \log(1 + e^x).$$

(Note that the derivative of LP(x) is the logistic function, hence the name "logistic potential.")

We may express (25) as

$$\log p(x) = \alpha + (x-y)^t\beta + LP(\gamma + (x-y)^t\delta). \qquad (26)$$

Thus, the Dutch Identity reveals that the two-class latent class model for dichotomous data is a log-nonlinear model of a very special form, (26). Different choices of y can affect α and γ in (26) but not β_j and δ_j. This representation of the two-point latent class model may yield alternative ways of fitting such models, and approximations to LP(X) may also prove useful.

3.5 What Does An Observed Response Vector Tell Us About The Value Of A Latent Variable?

The estimation of θ in IRT models is problematic. The LOGIST program, Wingersky (1983), produces "maximum likelihood" estimates of θ, $\hat{\theta}$, while the approach used in BILOG, Mislevy and Bock (1982), produces posterior expectations of θ given each possible response vector, y, i.e., $E(\theta|X=y)$. However, in my opinion, it has always been a mystery as to exactly what these quantities really mean since

a) the scale of θ is arbitrary,

b) for some choices of F, $E(\theta|X=y)$ will not exist,

c) the "likelihood function" used in LOGIST to compute $\hat{\theta}$'s is not the real likelihood function for many applications -- e.g., when examinees are sampled from a well-defined population the likelihood function in (2) is the correct one.

The Dutch Identity provides a key to understanding this mystery. The equation

$$\frac{p(x)}{p(y)} = E(e^{(x-y)^t\lambda}|X=y) \tag{27}$$

may be re-expressed in the following way. Let $r = x-y$ and let $S_y = \{r : y+r = x = \text{a 0/1 vector}\}$.

Thus S_y is the set of all (0,1,-1)-vectors r such that when added to y we get a (0,1)-vector, x, back. Clearly, S_y depends on y. Now (4) can be written as

$$E(e^{r^t\lambda}|X=y) = \frac{p(y+r)}{p(y)} \qquad \text{all } r\epsilon S_y. \tag{28}$$

Hence (28) says that for each fixed value of y, the moment generating function for the conditional distribution of λ given that X=y evaluated at each $r \epsilon S_y$ equals the ratio $p(y+r)/p(y)$. Since the manifest probabilities are, in principle, the most that the data can ever determine, equation (28) implies that for each y, the values of $E(e^{r^t\lambda}|X=y)$ for $r \epsilon S_y$ are the most that we can know about θ. Suppose we let

$$g_r(\lambda) = e^{r^t\lambda} , \tag{29}$$

then (28) says that

$$E(g_r(\lambda)|X=y) = \frac{p(y+r)}{p(y)} \tag{30}$$

for all $r \epsilon S_y$. Thus, (30) is an example of the so-called generalized moment problem. We are interested in the conditional distribution of λ given X=y. Equation (30) says that all we can know, in principle, are the values of the expectation of $g_r(\lambda)$ for all $r \epsilon S_y$ for this conditional distribution. Kemperman (1968) shows how knowledge of these generalized moments can be used to infer knowledge of the conditional distribution of λ given X=y. These inferences consist of <u>bounds</u> on probabilities of the form

$$L_y(S) \leq P(\gamma \epsilon S|X=y) \leq U_y(S), \tag{31}$$

where S is a set of λ-values.

Hence the mystery of what can be "estimated" about θ is resolved into bounds on probabilities of events that involve $\lambda(\theta)$ rather than θ. The tools developed by Kemperman (1968) and others can be used to investigate these issues further. The central role of $\lambda(\theta)$ in (30) suggests that model building ought to be in terms of $\lambda_j(\theta)$ rather than $P_j(\theta)$.

3.6 The Rasch Model

The Rasch model has a one-parameter logistic item response function which implies that the logit function, $\lambda_j(\theta)$, has the following linear form:

$$\lambda_j(\theta) = a(\theta - b_j). \tag{32}$$

In addition, the ability distribution, $F(\theta)$, is unspecified in the Rasch model. In Cressie and Holland (1983) it is shown that for the Rasch model the manifest probabilities, $\{p(x)\}$, have the following log-linear form:

$$\log p(x) = \alpha + \sum_j x_j \beta_j + \sum_k \gamma_k \delta(x_+, k) \tag{33}$$

where $x_+ = \Sigma x_j$ and

$$\delta(x_+, k) = \begin{cases} 1 & \text{if } x_+ = k \\ 0 & \text{otherwise.} \end{cases}$$

The parameters $\{\beta_j\}$ are unconstrained and each β_j may vary over $(-\infty, \infty)$. The $\{\gamma_k\}$ are the logarithms of a moment sequence, i.e.

$$\gamma_k = \log[E(U^k)],$$

for an arbitrary positive random variable U. Thus, the γ_k are subject to a system of inequalities described in detail in Cressie and Holland (1983) and in de Leeuw and Verhelst (1986).

The main tool used by Cressie and Holland to establish (33) is a form of the Dutch Identity with y=0. We may obtain an alternative formulation of (33) using the general Dutch Identity. This is given in Theorem 2.

Theorem 2: If p(x) satisfies an IRT model with one-parameter logistic IRFs (i.e., (32)) and general F, then p(x) satisfies the log-linear model

$$\log p(x) = \alpha + \sum_j (x_j - y_j)\beta_j + \gamma_k \delta(x_+, k), \tag{34}$$

where the β_j vary over $(-\infty, \infty)$ and the γ_k have the form

$$\gamma_k = \log[(E(U^{k-y_+})]$$

for $k = 0, 1, \ldots, J$, and U is an arbitrary positive random variable.

Proof: From the Dutch Identity we have

$$p(x) = p(y)\, E[\exp\{\sum_j (x_j - y_j)a(\theta - b_j)\} | X=y]$$

$$= p(y)\, \exp\{\sum_j (-ab_j)(x_j - y_j)\}\, E[\exp\{a\theta(x_+ - y_+)\} | X=y].$$

Now let $U = e^{a\theta}$ and take logs. This yields

$$\log p(x) = \alpha + \sum_j (x_j - y_j)\beta_j + \log E[U^{x_+ - y_+} | X=y]$$

where

$$\alpha = \log p(y), \text{ and } \beta_j = -ab_j.$$

Then set

$$\gamma_k = \log E[U^{k - y_+} | X=y]. \quad \text{QED}.$$

Cressie and Holland show that the total number of nonredundant parameters in (33) is 2J-1 -- there a J β's and J-1 γ's. However, the γ's are not freely varying parameters and are subject to a system of inequalities. While these inequalities do not restrict the γ's in a functional way, they do have an interesting impact on the values that the γ's can take on as the next corollary shows.

Corollary 2: If $p(x)$ satisfies the hypothesis of Theorem 2 and if y is such that $F(\theta | X=y)$ is $N(\mu_y, \sigma_y^2)$ then

$$\gamma_k = a\, \mu_y (k - y_+) + \tfrac{1}{2}(a\, \sigma_y)^2 (k - y_+)^2 \qquad (35)$$

so that the γ_k lie on a quadratic curve as a function of k.

Proof: Simply evaluate $E[e^{a\,\theta(x_+ - y_+)} | X=y]$ from the proof of Theorem 2 using the moment generating function of an univariate normal distribution. QED.

We observe that the existence of a y for which $\theta | X=y$ has a nearly normal distribution follows from the assumption that J is large and $F(\theta)$ is smooth.

Hence, rather than 2J-1 parameters, when J is large, the Rasch model can be expected to behave as though there were only J+1 parameters -- J β's and one γ (the coefficient of $(x_+-y_+)^2$).

4. DISCUSSION

In my opinion, the Dutch Identity has been shown to be a useful tool for the analysis of IRT models. The very fact it implies that long tests must exhibit very little third and higher-order interactions in their manifest probabilities, $\{p(\mathbf{x})\}$, is remarkable and not well-known. I have begun Monte Carlo simulation work to investigate how large J must be in order for SOE models to fit data generated by a model of the form (1). For ten items with Rasch IRFs the fit based on 30,000 simulated examinees is quite good -- likelihood ratio chi-squares of 965 on 968 degrees of freedom. For non-Rasch IRFs (either linear logit functions with different slopes or 3PL IRFs) the fit on ten items is not as good. These results are in agreement with the theory in this paper, but there is clearly more work to be done.

A second remarkable fact that the Dutch Identity implies concerns the number of parameters that can be estimated in a long test. The discussion in section 3.2 shows that all "smooth" unidimensional IRT models converge to a model of the form (19) as the number of items grows. The model in (19) has only two parameters per item which may be interpreted as the value of λ_j and of its first θ-derivative at a single point. Hence models that attempt to fit three or more parameters per item can only do so successfully for two reasons; either (1) they are not applied to a large enough item set or (2) the test is not unidimensional. What I conjecture from this analysis is that there is a sort of "conservation law" for IRT item parameters of the form: a D-dimensional set of J items can only support a total of (D+1)J parameters when J is large.

Individual items may be able to have more than D+1 parameters estimated for them, but only at the expense of fewer estimable parameters for some other items. The total cannot exceed (D+1)J. It will be very useful to see how this type of result actually works on real and simulated data.

REFERENCES

Birnbaum, A. (1968) Some latent trait models and their use in inferring an examinee's ability, in Lord, F. and Novick, M. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

Bishop, Y., Fienberg, S., and Holland, P. (1975) Discrete multivariate analysis: Theory and practice. Cambridge, MA: MIT Press.

Bock, R. D., and Aitken, M. (1981) Marginal maximum likelihood estimation of item parameters: application of the EM algorithm. Psychometrika, 46, 443-459.

Bock, R. D., and Lieberman, M. (1970) Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-197.

Bock, R. D. and Mislevy, R. (1982) Adaptive EAP estimation of ability in a micro-computer environment. Applied Psychological Measurement, 6, 431-444.

Cressie, N. (1982) "A useful empirical Bayes identity." Annals of Statistics, 10, 625-629.

Cressie, N. and Holland, P.W. (1983) "Characterizing the manifest probabilities of latent trait models." Psychometrika, 48, 129-141.

de Leeuw, J. and Verhelst, N. (1986) "Maximum likelihood estimation in generalized Rasch models." Journal of Educational Statistics, 11, 183-196.

Holland, P. (1981) When are item response theory models consistent with observed data? Psychometrika, 46, 79-92.

Holland, P., and Rosenbaum, P. (1986) Conditional association and unidimensionality in monotone latent variable models. Annals of Statistics, 14, 1523-1543.

Kemperman, J.H.B. (1968) "The general moment problem, a geometric approach." Annals of Mathematical Statistics, 39, 93-122.

Lord, F. (1962) Review of Studies in Item Analysis and Prediction (H. Solomon, ed.). Stanford, CA: Stanford University Press, 1961, in Psychometrika, 27, 207-213.

Mislevy, R. and Bock, R. D. (1982) BILOG: Item analysis and test scoring with binary logistic models [computer program]. Mooresville, IN: Scientific Software.

Tsao, Rhett (1967) "A second order exponential model for multidimensional dichotomous contingency tables with applications in medical diagnosis." Doctoral thesis, Department of Statisitics, Harvard University.

Wingersky, M. (1983) LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (ed.) Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.